Received: 9 March 2020 | Revised: 6 August 2020 | Accepted: 11 September 2020

DOI: 10.1002/sim.8766

Check for updates

RESEARCH ARTICLE

Statistics in Medicine WILEY

Minimum sample size for external validation of a clinical prediction model with a continuous outcome

Lucinda Archer[1] | **Kym I. E. Snell**[1] | **Joie Ensor**[1] | **Mohammed T. Hudda**[2] | **Gary S. Collins**[3] | **Richard D. Riley**[1]

[1]Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

[2]Population Health Research Institute, St George's, University of London, London, UK

[3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Correspondence
Lucinda Archer Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, Keele ST5 5BG, UK.
Email: l.archer@keele.ac.uk

Funding information
British Heart Foundation, Grant/Award Number: FS/17/76/33286; Cancer Research UK, Grant/Award Number: C49297/A27294; European Horizon 2020 research and innovation programme, Grant/Award Number: 777090; Medical Research Council; NIHR Biomedical Research Centre, Oxford; NIHR Clinical Trials Unit Support Funding; NIHR SPCR; NIHR SPCR Evidence Synthesis Working Group, Grant/Award Number: 390; Wellcome, Grant/Award Number: 102215/2/13/2

Clinical prediction models provide individualized outcome predictions to inform patient counseling and clinical decision making. External validation is the process of examining a prediction model's performance in data independent to that used for model development. Current external validation studies often suffer from small sample sizes, and subsequently imprecise estimates of a model's predictive performance. To address this, we propose how to determine the minimum sample size needed for external validation of a clinical prediction model with a continuous outcome. Four criteria are proposed, that target precise estimates of (i) R^2 (the proportion of variance explained), (ii) calibration-in-the-large (agreement between predicted and observed outcome values on average), (iii) calibration slope (agreement between predicted and observed values across the range of predicted values), and (iv) the variance of observed outcome values. Closed-form sample size solutions are derived for each criterion, which require the user to specify anticipated values of the model's performance (in particular R^2) and the outcome variance in the external validation dataset. A sensible starting point is to base values on those for the model development study, as obtained from the publication or study authors. The largest sample size required to meet all four criteria is the recommended minimum sample size needed in the external validation dataset. The calculations can also be applied to estimate expected precision when an existing dataset with a fixed sample size is available, to help gauge if it is adequate. We illustrate the proposed methods on a case-study predicting fat-free mass in children.

KEYWORDS
calibration, continuous outcomes, external validation, prediction model, sample size, R-squared

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2020 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Clinical prediction models provide individualized outcome predictions to inform patient counseling and clinical decision making, such as treatment and monitoring strategies.[1-3] Depending on the context, they may also be referred to as clinical prediction tools, diagnostic or prognostic models, risk scores, and prognostic indices, among other names. They are typically developed using a regression framework, which provides an equation to predict the outcome conditional on the values of multiple predictors (variables, covariates). In this article, we focus on prediction of continuous outcomes (such as birth weight, depression score, blood pressure or fat mass), for which the model equation is typically a linear regression. Such models can be used to predict an individual's expected outcome value, conditional on the individual's predictor values. The outcome may relate to something current (eg, fat mass level at present) or in the future (eg, pain score at 1 month after a back injury).

Recently we proposed how to calculate the minimum sample size needed to develop a prediction model with a continuous outcome.[4,5] Once a model has been developed, it is important to evaluate its predictive performance in new data, independent to that used to develop the model. This process is known as external validation, and is usually crucial regardless of how a model was developed. In particular, external validation indicates how the model performs in new data that is representative of the target population to which the model will be used in practice.[6-13] However, despite being widely encouraged and having its importance clearly demonstrated,[13-19] external validation of published prediction models is rare in practice, with researchers predominately focusing on the development of new models.[19] Even when external validation is performed, the sample size is often too small to provide reliable conclusions about a model's predictive performance and key measures are often neglected; in particular, calibration of predicted and observed outcome values is rarely examined.[16]

In this article, we propose criteria to determine the minimum sample size needed for external validation of a clinical prediction model with a continuous outcome. We suggest the minimum sample size needs to be large enough to precisely estimate three key measures of predictive performance: calibration slope (agreement between predicted and observed values across the range of predicted values), calibration-in-the-large (CITL, agreement between predicted and observed outcome values on average), and R^2 (the proportion of variance explained). Section 2 introduces these performance measures, while in Section 3, we derive three closed-form solutions for the sample size required to estimate each of them precisely. As these solutions depend on the variance of observed outcome values, we also present a fourth criterion that aims to ensure this variance is estimated precisely. Hence, our sample size calculation comprises checking four criteria, and we suggest the largest sample size calculated from the four approaches is used as the minimum required for the external validation. Section 4 applies our proposal to an applied example, and Section 5 concludes with discussion.

2 | KEY MEASURES OF PREDICTIVE PERFORMANCE FOR A CLINICAL PREDICTION MODEL WITH A CONTINUOUS OUTCOME

Assume that we wish to externally validate an existing prediction model for a continuous outcome, and have obtained a suitable external validation dataset containing a sample of individuals from the target population of interest. We now describe how to quantify the prediction model's performance in this dataset.

First, the researcher needs to calculate the existing model's predicted (expected) outcome value ($Y_{\text{PRED}i}$) for each individual (i). As the outcome is continuous, the existing prediction model equation will usually be in the form of a linear regression and so contain an intercept (α), and predictor effects (β_1, β_2, β_3, etc) corresponding to predictor variables (X_{1i}, X_{2i}, X_{3i}, etc). For example, with three predictors a simple example of an existing prediction model equation is:

$$Y_{\text{PRED}i} = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}. \quad (1)$$

However, in practice the right hand side of the model equation (also known as the model's *linear predictor*) may be far more complex, for instance with more than three predictors and potential interactions and non-linear terms (eg, defined by splines or polynomials). A real example is given in Box 1.

BOX 1 Hudda et al prediction model for the natural logarithm of ln(fat-free mass) in children[20]

$$\text{Ln(fat-free mass)} = 2.8055 + (0.3073 \times \text{height}^2) - (10.0155 \times \text{weight}^{-1}) + (0.004571 \times \text{weight}) + (0.01408 \times \text{BA}) - (0.06509 \times \text{SA}) - (0.02624 \times \text{AO}) - (0.01745 \times \text{other}) - (0.9180 \times \ln(\text{age})) + (0.6488 \times \text{age}^{0.5}) + (0.04723 \times \text{male})$$

- Predictor variables of black (BA), south Asian (SA), other Asian (AO), or other (other) ethnic origins are all binary, with value of 1 if individual has the particular origin and 0 otherwise
- Height is measured in meters, weight in kilograms, age in years, and fat-free mass in kilograms

Clearly, the external validation dataset must contain values for all the predictors (X_{1i}, X_{2i}, X_{3i}, …) included in the prediction model equation, so that $Y_{\text{PRED}i}$ can be calculated by applying the model's equation to each individual. The dataset must also contain the observed outcome value (Y_i) for each individual, so that the prediction model's predictive performance can then be quantified by comparing the $Y_{\text{PRED}i}$ values to the Y_i values.

We now introduce three key statistics to quantify a model's predictive performance upon external validation, which focus on overall model fit and calibration.

2.1 | R-squared

R^2 is a well-known measure of overall model fit and quantifies the proportion of outcome variation explained.

Let $\text{var}(Y_i)$ denote the variance of Y_i values in the external validation population, and $\text{var}(Y_i - Y_{\text{PRED}i})$ denote the variance of $(Y_i - Y_{\text{PRED}i})$ values (ie, the prediction errors in the external validation population). Then the true proportion of outcome variation explained by the predicted values from the prediction model, R^2_{val}, is:

$$R^2_{\text{val}} = 1 - \left(\frac{\text{var}(Y_i - Y_{\text{PRED}i})}{\text{var}(Y_i)} \right). \tag{2}$$

Values of R^2_{val} closer to 1 indicate better fit of the $Y_{\text{PRED}i}$ from the prediction model.

2.2 | Calibration slope and calibration-in-the-large

Calibration measures the agreement between predicted ($Y_{\text{PRED}i}$) and observed (Y_i) outcome values in the external validation dataset.[21] It is best shown graphically on a calibration plot, with $Y_{\text{PRED}i}$ on the horizontal axis plotted against Y_i on the vertical axis, with every individual providing a single data point. A LOESS smoothed calibration curve should also be fitted through the points and presented on the plot.[2,11,22] Ideally, the predicted outcome values are not systematically under- or over-estimated across the entire range of predicted values. That is, the points are scattered randomly around the 45° line of perfect agreement (corresponding to a slope of 1), with little variation around the line (ie, $\widehat{R}^2_{\text{val}}$ is large), and with close agreement between predicted and observed values across the entire horizontal axis range.

To formally quantify calibration performance in an external validation dataset, a calibration model can be fitted of the form,

$$Y_i = \alpha_{\text{cal}} + \lambda_{\text{cal}}(Y_{\text{PRED}i}) + e_{\text{cal}i}$$

$$e_{\text{cal}i} \sim N(0, \sigma^2_{\text{cal}}), \tag{3}$$

where "cal" is used to emphasize that parameters are from the calibration model. This model can be fitted using standard estimation methods for a linear regression, such as using restricted maximum likelihood estimation. The parameter λ_{cal} represents the *calibration slope*, which measures agreement between predicted and observed outcomes across the whole range of predicted values.[2,3] As mentioned, the ideal λ_{cal} value is 1. A $\lambda_{\text{cal}} < 1$ indicates that some predictions are too

extreme (eg, predictions above the mean are too high, and/or predictions below the mean are too low) and a slope > 1 indicates that the range of predictions is too narrow. A calibration slope < 1 is often observed in external validation studies, as clinical prediction models are often developed in small datasets without adjustment for overfitting, which leads to extreme predictions (miscalibration) in new individuals external to those used for model development.[23-26] The term σ^2_{cal} measures the residual variance in the calibration model.

Note that the true calibration slope in the external validation population can also be expressed as,[27]

$$\lambda_{\text{cal}} = \sqrt{\frac{R^2_{\text{cal}}\text{var}(Y_i)}{\text{var}(Y_{\text{PRED}i})}}, \tag{4}$$

where R^2_{cal} is the proportion of variance of Y_i values explained when the calibration model (3) is fitted to the external validation population.

Systematic over- or under-prediction is still possible even when the calibration slope is 1, and thus it should always be considered alongside calibration plots and CITL. The latter measures the agreement between mean predicted ($\overline{Y}_{\text{PRED}}$) and mean observed ($\overline{Y}$) outcome values, which can be estimated in the external validation dataset using:

$$\widehat{\text{CITL}}_{\text{val}} = \overline{Y} - \overline{Y}_{\text{PRED}}. \tag{5}$$

Estimating CITL_{val} by applying Equation (5) in an external validation dataset is equivalent to estimating α_{cal} by fitting model (3) with the constraint that λ_{cal} equals 1 (see Section 3.2).

3 | SAMPLE SIZE REQUIRED TO TARGET PRECISE ESTIMATES OF PREDICTIVE PERFORMANCE

In this section, we propose four criteria for researchers to use as a basis for determining the minimum sample size required for an external validation study. The first three criteria aim to ensure the sample size is large enough to estimate R^2_{val}, CITL_{val}, and λ_{cal} precisely (ie, with a small margin of error). Closed-form solutions are derived for this purpose. As these expressions depend on the estimates of (residual) variances, a fourth criterion aims to precisely estimate these also.

3.1 | Criterion (i): Precise estimate of R^2_{val}

Our first criterion targets a precise estimate for R^2_{val} from the external validation dataset, such that the confidence interval for R^2_{val} will be narrow. There are many suggestions for deriving confidence intervals for R^2.[28] Here, we focus on the approach suggested by Wishart,[29] which uses the following approximate standard error (SE) of $\widehat{R}^2_{\text{val}}$:

$$\text{SE}_{\widehat{R}^2_{\text{val}}} = \sqrt{\frac{4R^2_{\text{val}}(1-R^2_{\text{val}})^2}{n}}. \tag{6}$$

Tan suggests this approximation works well when the sample size (n) is reasonably large (say >50),[28] which is likely to be the situation when externally validating a clinical prediction model (see criterion (iv)). Rearranging Equation (6) gives a closed-form sample size calculation of:

$$n = \frac{4R^2_{\text{val}}(1-R^2_{\text{val}})^2}{\text{SE}^2_{\widehat{R}^2_{\text{val}}}}. \tag{7}$$

Equation (7) can now be used to calculate the sample size (n) required to meet criterion (i), by specifying a desired value for $\text{SE}_{\widehat{R}^2_{\text{val}}}$ and by setting R^2_{val} at the anticipated true value for the external validation population.

For example, consider an existing prediction model with an adjusted R^2 of 0.5 in the development dataset, with this adjusted (rather than apparent) R^2 giving an unbiased estimate of expected performance in new data. Then, if we assume

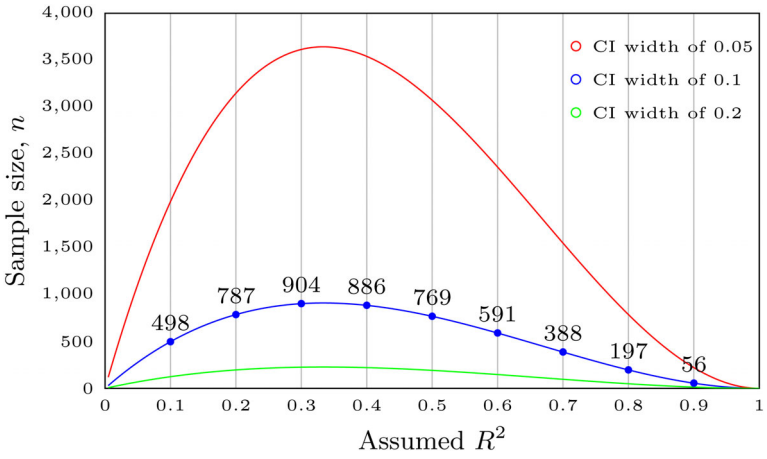

FIGURE 1 Sample size (number of participants, n) needed in an external validation dataset to target a confidence interval for R^2_{val} of a particular width (either 0.05, 0.1, or 0.2) for different assumed R^2_{val} values between 0.1 and 0.9. Sample size calculated using Equation (7) [Color figure can be viewed at wileyonlinelibrary.com]

the validation sample is from a similar target population to the development sample, a simple starting point is to anticipate R^2_{val} upon external validation is similar to the adjusted $\widehat{R}^2$ reported in the model development study. To target a 95% confidence interval for R^2_{val} that has a narrow width of about 0.1, we need a small $\text{SE}_{\widehat{R}^2_{\text{val}}}$ of 0.0255. This stems from assuming a 95% confidence interval for R^2_{val} can be derived approximately by $\widehat{R}^2_{\text{val}} \pm (1.96 \times \text{SE}_{\widehat{R}^2_{\text{val}}})$. We can now apply Equation (7) to give,

$$n = \frac{4R^2_{\text{val}}(1-R^2_{\text{val}})^2}{\text{SE}^2_{\widehat{R}^2_{\text{val}}}} = \frac{4 \times 0.5 \times (1-0.5)^2}{0.0255^2} = 768.9$$

and so 769 participants are required to meet criterion (i). To achieve the same margin of error, 905 participants are required when assuming R^2_{val} is 0.3, and 197 participants are required when assuming R^2_{val} is 0.8. These values are reasonably close to those using more exact (but not closed-form) approaches to confidence interval derivation, such as that based on the scaled non-central F approximation proposed by Lee.[30] The *ss.aipe.R2* function within Kelley's MBESS package for the R software identifies the sample size required to ensure Lee's confidence interval for R^2_{val} is sufficiently narrow,[31-33] and so is an alternative to using Equation (7).

Figure 1 shows how the required sample size changes from R^2_{val} values between 0.1 and 0.9 based on Equation (7) and assuming $\text{SE}_{\widehat{R}^2_{\text{val}}}$ is 0.0255 to target a confidence interval width of 0.1. The required sample size will be lower when allowing for wider target confidence intervals, and higher when aiming for narrower target confidence intervals (Figure 1). However, we suggest $\text{SE}_{\widehat{R}^2_{\text{val}}} \leq 0.0255$ is a sensible compromise, as it targets a precise estimate (margin of error of 0.05 or less compared to the true value) and still gives a required sample size that will be realistic to obtain in practice.

Note that upon external validation the true R^2_{val} may be lower or higher than the adjusted $\widehat{R}^2$ reported for model development. Therefore, although the adjusted $\widehat{R}^2$ from the development study is a useful starting point, we also recommend calculating the sample size required when assuming slightly different values for the true R^2_{val}. For example, researchers might apply Equation (7) assuming R^2_{val} values $\pm$ 0.1 of the adjusted $\widehat{R}^2$ reported from the development study, and note the largest sample size across this range.

3.2 | Criterion (ii): Precise estimate of CITL

Our second criterion targets a precise estimate of CITL_{val} from the external validation dataset. We estimate CITL_{val} by using $\overline{Y} - \overline{Y}_{\text{PRED}}$ (from Equation (5)), which is equivalent to estimating the intercept when fitting (in the external validation dataset) model (3) with the predicted values as an offset term:

$$Y_i = \text{CITL}_{\text{val}} + 1(Y_{\text{PRED}i}) + e_{\text{CITL}i}$$

$$e_{\text{CITL}i} \sim N(0, \sigma^2_{\text{CITL}}). \quad (8)$$

Therefore the SE of $\widehat{\text{CITL}}$ is:

$$\text{SE}^2_{\widehat{\text{CITL}}} = \text{var}(\overline{Y} - \overline{Y}_{\text{PRED}}) = \text{var}\left(\frac{\sum_{i=1}^{n}(Y_i - Y_{\text{PRED}i})}{n}\right) = \frac{\sigma^2_{\text{CITL}}}{n} = \frac{\text{var}(Y_i)(1 - R^2_{\text{CITL}})}{n}. \quad (9)$$

We can rearrange Equation (9) to obtain an expression for the required sample size:

$$n = \frac{\text{var}(Y_i)(1 - R^2_{\text{CITL}})}{\text{SE}^2_{\widehat{\text{CITL}}}}. \quad (10)$$

Hence, the sample size required to meet criterion (ii) can be derived using Equation (10), for which the researcher must pre-specify R^2_{CITL} (the anticipated proportion of variance explained by the predictions in the external validation population), along with $\text{var}(Y_i)$ (the anticipated variance of Y_i in the target population), and the desired $\text{SE}_{\widehat{\text{CITL}}}$.

A sensible starting point is to assume CITL is zero, as then $R^2_{\text{CITL}} = R^2_{\text{val}}$ (the anticipated proportion of variance explained by the predictions upon validation), and so

$$n = \frac{\text{var}(Y_i)(1 - R^2_{\text{val}})}{\text{SE}^2_{\widehat{\text{CITL}}}}, \quad (11)$$

with R^2_{val} assumed to be the same as the adjusted $\widehat{R}^2$ reported from the development study.

If CITL is not zero then R^2_{CITL} will not equal R^2_{val}. Hence, it is also sensible to consider a range of values for R^2_{CITL} when applying Equation (10), such as $\pm$ 0.1 of the adjusted $\widehat{R}^2$ reported from the development study, and to note the largest sample size across this range.

The value that defines a precise $\text{SE}_{\widehat{\text{CITL}}}$ is context specific, as it depends on the scale of the outcome values. For example, for systolic blood pressure an SE of about 2.5 mmHg may suffice, but for BMI a smaller SE may be required as the scale is much narrower.

For instance, consider external validation of a prediction model for systolic blood pressure with a reported adjusted R^2 of 0.5 in the development study, and that the variance of the observed Y_i values is anticipated to be 400 in the target population for the validation study. Let us target an $\text{SE}_{\widehat{\text{CITL}}}$ of 2.55, as this gives a 95% confidence interval for CITL_{val} with a narrow width of about 10 mmHg, assuming a 95% confidence interval for CITL_{val} can be derived approximately by $\widehat{\text{CITL}} \pm (1.96 \times \text{SE}_{\widehat{R}^2_{\text{val}}})$. Assuming $R^2_{\text{CITL}} = R^2_{\text{val}} = 0.5$, then applying Equation (10) gives,

$$n = \frac{\text{var}(Y_i)(1 - R^2_{\text{val}})}{\text{SE}^2_{\widehat{\text{CITL}}}} = \frac{400 \times (1 - 0.5)}{2.55^2} = 30.76$$

and thus at least 31 participants are required to achieve criterion (ii).

More cautiously assuming that $R^2_{\text{CITL}} = 0.4$, the required sample size is

$$n = \frac{\text{var}(Y_i)(1 - R^2_{\text{CITL}})}{\text{SE}^2_{\widehat{\text{CITL}}}} = \frac{400 \times (1 - 0.4)}{2.55^2} = 36.91$$

and thus 37 participants are required.

It is likely that the sample size to precisely estimate CITL is smaller than that required to precisely estimate the measures outlined in criteria (i), (iii), and (iv).

3.3 | Criterion (iii): Precise estimate of calibration slope

The third criterion targets a precise estimate of λ_{cal}, which represents the calibration slope obtained from fitting calibration model (3) in the external validation dataset. As $\widehat{\lambda}_{\text{cal}}$ is the slope from a simple linear regression model, the SE of $\widehat{\lambda}_{\text{cal}}$ can be estimated by,[34]

$$\mathrm{SE}^2_{\hat{\lambda}_{\mathrm{cal}}} = \frac{\sigma^2_{\mathrm{cal}}}{\sum_{i=1}^{n}(Y_{\mathrm{PREDi}} - \overline{Y}_{\mathrm{PRED}})^2},$$

where σ^2_{cal} is the residual variance from model (3).

By utilizing Equation (4), and also recognizing that $\sigma^2_{\mathrm{cal}} = \mathrm{var}(Y_i)(1 - R^2_{\mathrm{cal}})$ and that $\sum_{i=1}^{n}(Y_{\mathrm{PRED}i} - \overline{Y}_{\mathrm{PRED}})^2 = (n-1)\,\mathrm{var}(Y_{\mathrm{PRED}i})$, we can write $\mathrm{SE}^2_{\hat{\lambda}_{\mathrm{cal}}}$ in terms of λ^2_{cal} and R^2_{cal} values,[27] as follows:

$$\begin{aligned}
\mathrm{SE}^2_{\hat{\lambda}_{\mathrm{cal}}} &= \frac{\sigma^2_{\mathrm{cal}}}{\sum_{i=1}^{n}(Y_{\mathrm{PRED}i} - \overline{Y}_{\mathrm{PRED}})^2} \\
&= \frac{\mathrm{var}(Y_i)(1 - R^2_{\mathrm{cal}})}{(n-1)\,\mathrm{var}(Y_{\mathrm{PRED}i})} \\
&= \frac{\mathrm{var}(Y_i)}{(n-1)\,\mathrm{var}(Y_{\mathrm{PRED}i})} - \frac{\mathrm{var}(Y_i)R^2_{\mathrm{cal}}}{(n-1)\,\mathrm{var}(Y_{\mathrm{PRED}i})} \\
&= \frac{\mathrm{var}(Y_i)}{(n-1)\,\mathrm{var}(Y_{\mathrm{PRED}i})} - \frac{\lambda^2_{\mathrm{cal}}}{(n-1)} \\
&= \frac{\lambda^2_{\mathrm{cal}}}{(n-1)\,R^2_{\mathrm{cal}}} - \frac{\lambda^2_{\mathrm{cal}}R^2_{\mathrm{cal}}}{(n-1)R^2_{\mathrm{cal}}} \\
&= \frac{\lambda^2_{\mathrm{cal}}(1 - R^2_{\mathrm{cal}})}{(n-1)R^2_{\mathrm{cal}}}.
\end{aligned} \tag{12}$$

Rearranging gives:

$$n = \frac{\lambda^2_{\mathrm{cal}}(1 - R^2_{\mathrm{cal}})}{\mathrm{SE}^2_{\hat{\lambda}_{\mathrm{cal}}}\,R^2_{\mathrm{cal}}} + 1. \tag{13}$$

Equation (13) allows calculation of the required sample size for a desired $\mathrm{SE}_{\hat{\lambda}_{\mathrm{cal}}}$, conditional on specifying λ_{cal} (the anticipated (mis)calibration across the range of predicted values) and R^2_{cal} (the anticipated proportion of variance in observed Y_i values explained by the calibration model).

In terms of choosing $\mathrm{SE}^2_{\hat{\lambda}_{\mathrm{cal}}}$, a value ≤ 0.051 is recommended, to target a 95% confidence interval for λ_{cal} that has a narrow width ≤ 0.2 (eg, if the calibration slope was 1, the confidence interval would be 0.9 to 1.1 assuming confidence intervals derived by $\hat{\lambda}_{\mathrm{cal}} \pm 1.96\mathrm{SE}_{\hat{\lambda}_{\mathrm{cal}}}$; note that replacing 1.96 by critical values of the t-distribution is unnecessary, as the sample size will not be small).

In terms of choosing λ_{cal}, a simple starting point is to assume good calibration, such that $\lambda_{\mathrm{cal}} = 1$ and $\alpha_{\mathrm{cal}} = 0$ in model (3). Then, $R^2_{\mathrm{cal}} = R^2_{\mathrm{val}}$ from criterion (i), and so R^2_{cal} might be assumed to be the same as the adjusted R^2 estimated in the model development study. For example, for external validation of a prediction model that had an estimated adjusted R^2 of 0.5 in the development dataset, a simple starting point is to anticipate the same value for R^2_{val}. Then, assuming the model's predictions will be well calibrated in the external validation dataset (ie, on average, fitting model (3) would give $\hat{\alpha}_{\mathrm{cal}}$ of 0 and a $\hat{\lambda}_{\mathrm{cal}}$ of 1), using Equation (13) gives,

$$n = \frac{\lambda^2_{\mathrm{cal}}(1 - R^2_{\mathrm{cal}})}{\mathrm{SE}^2_{\hat{\lambda}_{\mathrm{cal}}}\,R^2_{\mathrm{cal}}} + 1 = \frac{1 \times (1 - 0.5)}{0.051 \times 0.051 \times 0.5} + 1 = 385.47$$

and thus 386 participants are required to target a confidence interval width of 0.1 for the calibration slope, under the assumptions of good calibration.

The sample size should also be large enough to precisely estimate some miscalibration. Often when a prediction model is externally validated the calibration slope is less than 1, due to overfitting during model development that was unaccounted for in the final prediction model equation (ie, penalization or shrinkage estimation methods were not used). In such situations R^2_{cal} can still be assumed to be the same as the adjusted R^2 presented for model development, as this

value specifically adjusts for optimism due to overfitting. When applying Equation (13) for fixed R^2_{cal} and $SE^2_{\hat{\lambda}_{cal}}$ values, lowering the assumed λ_{cal} below 1 will produce lower sample sizes than when assuming the prediction model is well calibrated. Hence, assuming λ_{cal} is 1 is more conservative for the sample size calculation.

Further sensitivity analyses could be undertaken if desired. For example, we could change both λ_{cal} and R^2_{cal} values. However this is complex, as Equation (4) reveals that the value of λ_{cal} depends on R^2_{cal} (and also var(Y_i) and var(Y_{PREDi})). Therefore, changing the assumed value of λ_{cal} has implications for what the assumed value of R^2_{cal} should be. This may be too intricate for the sample size calculation. Similarly, although situations of under-prediction (where λ_{cal} is >1) may lead to larger required sample sizes, this may not be practical to consider as over-prediction situations are more common. Thus, we generally suggest to apply Equation (13) assuming good calibration (λ_{cal} =1) and set R^2_{cal} equal to the adjusted R^2 estimated for model development.

3.4 | Criterion (iv): Precise estimates of residual variances

Our final criterion targets precise estimates of $\hat{\sigma}^2_{CITL}$ and $\hat{\sigma}^2_{cal}$. This is essential because, although these residual variances are not direct measures of predictive performance themselves, their estimated values are used toward parameter estimates and, crucially, $SE_{\widehat{CITL}_{val}}$ and $SE_{\hat{\lambda}_{cal}}$.

For $\hat{\sigma}^2_{CITL}$, we can equivalently consider the sample size needed to precisely estimate a residual variance in a linear regression model with only an intercept (see model (8)). In such situations, Harrell suggests calculating the sample size to ensure the lower and upper bounds of a 95% confidence interval for the residual variance has a small multiplicative margin of error (MMOE) around the true value,[2] using

$$\text{MMOE} = \sqrt{\max\left(\frac{\chi^2_{1-\frac{\alpha}{2},n-1}}{n-1}, \frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}\right)}, \tag{14}$$

where $\chi^2_{1-\frac{\alpha}{2},n-1}$ and $\chi^2_{\frac{\alpha}{2},n-1}$ are the critical values of a χ^2 distribution with $n-1$ degrees of freedom for which there is, respectively, a probability of $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ of being less than the critical value. The second term within the bracket of Equation (14) will typically give the largest MMOE.

We recommend a margin of error of within 10% of the true value ($1.0 \leq \text{MMOE} \leq 1.1$), for which Equation (14) reveals that a sample size of at least 234 participants is needed to ensure an MMOE ≤1.1 for $\hat{\sigma}^2_{CITL}$.

For precise estimation of $\hat{\sigma}^2_{cal}$, we need to adjust the sample size required for a slope parameter being estimated (see model (3)). As outlined by Riley et al,[4] the solution is simply 234 + 1, and thus 235 participants are required to ensure an MMOE of ≤1.1 for $\hat{\sigma}^2_{cal}$. Hence, in summary, at least 235 participants are needed to meet criterion (iv), and thus 235 is the minimum sample size required for any external validation of a prediction model for a continuous outcome, regardless of context and before consideration of criteria (i), (ii), or (iii).

3.5 | Summary of the criteria

Our sample size criteria aim to ensure the external validation dataset will precisely estimate R^2_{val}, CITL, calibration slope, and residual variances. The approach requires a separate sample size calculation for each criterion, and the largest sample size calculated provides the minimum needed for the external validation study. A step-by-step guide to our proposal is provided in Figure 2.

4 | APPLIED EXAMPLE

We now illustrate our sample size proposal using an applied example. Hudda et al developed a prediction model for the natural logarithm of fat-free mass in children and adolescents aged 4 to 15 years, including 10 predictor parameters based on height, weight, age, sex, and ethnicity (see Box 1 for model equation).[20] The model is required to provide an estimate of an individual's current fat mass (weight - predicted fat-free mass). The apparent calibration of the model in the development dataset is shown in Figure 3A. In the development dataset, the estimated adjusted R^2 was 0.95. An initial

STEP 1: Calculate the sample size needed to precisely estimate R^2_{val} (criterion (i))

Apply equation 7,

$$n = \frac{4R^2_{val}(1-R^2_{val})^2}{SE^2_{\hat{R}^2_{val}}}$$

after specifying suitable values for $SE_{\hat{R}^2_{val}}$ and R^2_{val}. We recommend using $SE_{\hat{R}^2_{val}} \leq 0.0255$ (to target a 0.1 width), and initially choosing R^2_{val} to equal the adjusted $\hat{R}^2$ reported for the model development study. A few other values for R^2_{val} might also be considered (e.g. values ±0.1 the adjusted $\hat{R}^2$ reported from the development study).

STEP 2: Calculate the sample size needed to precisely estimate calibration-in-the-large (criterion (ii))

Apply equation 10,

$$n = \frac{var(Y_i)(1-R^2_{CITL})}{SE^2_{\hat{CITL}}}$$

after specifying suitable values for R^2_{CITL} (akin to those used for R^2_{val} in step 1), $SE_{\hat{CITL}}$ and $var(Y_i)$. The latter represents the variance of outcome values in the population of interest, and should be based on other existing knowledge (e.g. previous studies). The value of $SE_{\hat{CITL}}$ should aim to ensure that $(\bar{Y} - \bar{Y}_{PRED}) \pm (1.96 * SE_{\hat{CITL}})$ is narrow, and so needs to be chosen in context of what constitutes a precise estimate of the anticipated mean prediction error in the clinical setting of interest.

STEP 3: Calculate the sample size needed to precisely estimate calibration slope (criterion (iii))

Apply equation 13,

$$n = \frac{\lambda^2_{cal}(1-R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1$$

after specifying suitable values for λ_{cal}, R^2_{cal} and $SE_{\hat{\lambda}_{cal}}$. We recommend $SE_{\hat{\lambda}_{cal}} \leq 0.051$ (to target a confidence interval width ≤ 0.2), choosing R^2_{cal} to be that chosen for R^2_{val} (i.e. the adjusted R^2 reported from the model development study; see step 1), and $\lambda_{cal} = 1$ (good calibration).

STEP 4: Calculate the sample size for precisely estimating residual variances (criterion (iv))

To target residual variance estimates in the calibration models that have a margin of error of $\leq 10\%$, at least 235 participants are required based on equation 14.

STEP 5: Calculate the final sample size

The required minimum sample size is the maximum value from steps 1 to 4, to ensure that each of criteria (i) to (iv) are met.

FIGURE 2 Summary of the steps involved in our sample size calculation for external validation of a clinical prediction model for a continuous outcome

external validation was undertaken in 176 children aged 11-12 years from the UK Avon Longitudinal Study of Parents and Children (ALSPAC) study,[35,36] where the model had an estimated R^2_{val} of 0.90 Figure 3B. However, as acknowledged by Hudda et al, further external validation is warranted in a broader age range, for which a sample size calculation can be undertaken using our proposal. We assume that the validation population is similar to the development population, and work through the calculations for criteria (i) to (iv).

STEP 1: Calculate the sample size needed to precisely estimate R^2_{val} (criterion (i))

This requires us to apply Equation (7). Based on assuming an $R^2_{\text{val}} = 0.90$, as in the published external validation of the model, and a $\text{SE}_{\hat{R}^2_{\text{val}}}$ of 0.0255 to target a confidence interval width of 0.1, a sample size of 56 children is required, as:

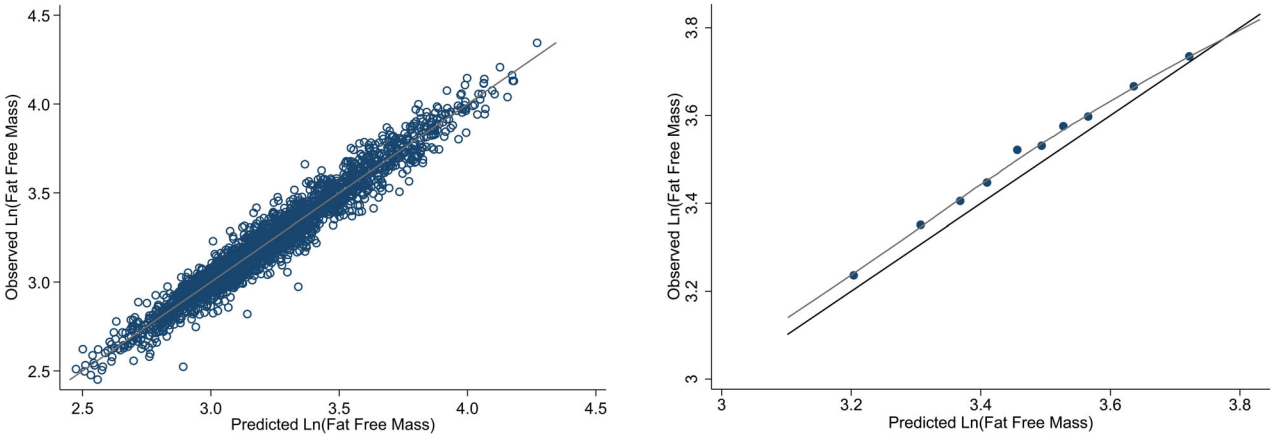

FIGURE 3 Calibration performance: A, in the development dataset; and, B, on external validation of the prediction model for ln(fat-free mass) in children, as proposed by Hudda et al.[20] The 45° line shows perfect calibration on both plots. * in B, individual level data points cannot be shown for confidentiality reasons. Data points shown are mean predicted against mean observed ln(fat-free mass) within tenths of predicted ln(fat-free mass), with a local regression smoother through the individual level data points shown in gray [Color figure can be viewed at wileyonlinelibrary.com]

$$n = \frac{4R^2_{\text{val}}(1-R^2_{\text{val}})^2}{\text{SE}^2_{\hat{R}^2_{\text{val}}}} = \frac{4 \times 0.90 \times (1-0.90)^2}{0.0255^2} = 55.4.$$

It is sensible to also consider that the model may perform worse upon external validation, say with a 0.1 reduction in R^2_{val} to 0.80. Then, the required sample size to meet criterion (i) is 197 children. These sample size values are also identified within Figure 1.

STEP 2: Calculate the sample size needed to precisely estimate calibration-in-the-large (criterion (ii))

This requires us to apply Equation (10), which itself requires us to specify var(Y_i), the anticipated variance of outcome values in the target population for external validation. Let us illustrate how to derive this from published information. In their paper, Hudda et al reported the lower quartile (LQ) as 20.8 and the upper quartile (UQ) as 30.6 kg of fat-free mass in their development dataset. By transforming this to the ln(kg) scale, and assuming ln(fat-free mass) values are approximately normally distributed, we can derive an estimate of the SD of the ln(fat-free mass) in the development population using[37]:

$$\frac{\ln UQ - \ln LQ}{1.35} = \frac{\ln 30.6 - \ln 20.8}{1.35} = 0.286.$$

Therefore, based on the published information $\widehat{\text{var}}(Y_i) \approx 0.286^2 = 0.082$. Interestingly, when contacting the original study authors directly for this information, they calculated it to be a similar value of $\widehat{\text{var}}(Y_i) = 0.089$. We will use this value from the study authors going forward.

We must also specify the expected value for R^2_{CITL}. We begin by assuming $R^2_{\text{CITL}} = R^2_{\text{val}}$ and that this is 0.90, as in Hudda's initial external validation of the model.

The precision required to estimate CITL needs to be placed in context of the mean outcome value in the population. Hudda et al reported a median baseline fat-free mass of 24.8 kg. If we assume that the mean value is similar, then we have:

$$\overline{Y} \approx \ln 24.8 = 3.21.$$

Considering the original untransformed scale, an accuracy of approximately ±1 kg around $\overline{Y}$ seems reasonably precise. A confidence interval of about 23.8 to 25.8 on the kg scale would correspond to a 95% CI of about 3.17 to 3.25 around $\overline{Y}$, implying a target $\text{SE}^2_{\widehat{\text{CITL}}}$ of about 0.02.

TABLE 1 Summary of the sample size calculation for external validation of the prediction model of Hudda et al

Criterion	Target precision	Assumptions	Minimum sample size required
(i) Precise estimate of $\boldsymbol{R^2_{val}}$	$SE_{\widehat{R^2_{val}}} = 0.0255$	$R^2_{val} = 0.8$	197
		$R^2_{val} = 0.9$	56
(ii) Precise estimate of **CITL**	$SE_{\widehat{CITL}} = 0.02$	$R^2_{CITL} = R^2_{val} = 0.8$, $var(Y_i) = 0.089$	45
		$R^2_{CITL} = R^2_{val} = 0.9$, $var(Y_i) = 0.089$	23
(iii) Precise estimate of $\boldsymbol{\lambda_{cal}}$	$SE_{\widehat{\lambda}_{cal}} = 0.051$	$R^2_{cal} = R^2_{val} = 0.9$ $\widehat{\lambda}_{cal} = 1$	44
(iv) Precise $\boldsymbol{\widehat{\sigma}^2_{CITL}}$ and $\boldsymbol{\widehat{\sigma}^2_{cal}}$	$1.0 \leq MMOE \leq 1.1$	-	235

Therefore, we can now apply Equation (10) to obtain a sample size of,

$$n = \frac{var(Y_i)\,(1 - R^2_{CITL})}{SE^2_{\widehat{CITL}}} = \frac{0.089 \times (1 - 0.9)}{0.02^2} = 22.3,$$

and thus 23 participants are required to meet criterion (ii). To be conservative, let us assume a 0.1 lower value for R^2_{CITL} to 0.80. Then, the required sample size to meet criterion (ii) would increase to 45 children.

STEP 3: Calculate the sample size needed to precisely estimate calibration slope (criterion (iii))

This requires us to apply Equation (13) after choosing values for $SE_{\widehat{\lambda}_{cal}}$, R^2_{cal}, and λ^2_{cal}. Let us choose an $SE_{\widehat{\lambda}_{cal}}$ of 0.051 to target a confidence interval width of 0.2. Further, we assume $R^2_{cal} = R^2_{val}$ and take the value of 0.90 as reported by the initial validation study of Hudda et al; and assume good calibration such that λ^2_{cal} is 1. We can now apply Equation (13) to give,

$$n = \frac{\lambda^2_{cal}(1 - R^2_{cal})}{SE^2_{\widehat{\lambda}_{cal}} R^2_{cal}} + 1 = \frac{1 \times (1 - 0.9)}{0.051^2 \times 0.9} + 1 = 43.72,$$

and thus 44 participants are required.

STEP 4: Calculate the sample size for precisely estimating residual variances (criterion (iv))

To ensure a 10% margin of error in residual variance estimates from the calibration models, at least 235 participants are required (see Section 3.4).

STEP 5: Calculate the final sample size

Assuming we aim to validate the model of Hudda et al in a population similar to the development data, steps 1 to 4 have provided four sample sizes to ensure criteria (i) to (iv) are met. These are summarized in Table 1. Based on the largest of these sample sizes, the final minimum sample size required to meet all criteria is 235 participants. This is driven by criterion (iv), to target sufficient precision around $\widehat{\sigma}^2_{CITL}$ and $\widehat{\sigma}^2_{cal}$.

5 | WHAT IF SAMPLE SIZE FOR EXTERNAL VALIDATION IS FIXED?

Sometimes there are no resources for prospective recruitment of participants to a new study for external validation of a prediction model. Then, researchers might seek an existing (already collected) dataset from the target population of interest. However, the sample size of an existing dataset is fixed, and so the researcher (and other stakeholders such as funders and collaborators) needs to know if it is large enough for reliable external validation. In this situation, our calculations in steps 1 to 4 can be re-expressed to calculate the expected $SE_{\widehat{R^2_{val}}}$, $SE_{\widehat{CITL}}$, $SE_{\widehat{\lambda}_{cal}}$, and MMOE conditional on the known sample size and assumed values of R^2_{val}, $var(Y_i)$, R^2_{CITL}, R^2_{cal}, and λ_{cal} as before.

For example, in the initial external validation of Hudda et al, an existing dataset, from the ALSPAC study, of 176 children was used. Based on the calculation shown in Table 1, this sample size is likely to give very precise estimates of R^2_{val}, CITL, and λ_{cal} when assuming $R^2_{val} = R^2_{CITL} = R^2_{cal}$ is 0.9. However, the sample size is lower than the 235 recommended for precise estimation of $\widehat{\sigma}^2_{CITL}$ and $\widehat{\sigma}^2_{cal}$, and so the MMOE for these estimates is expected to be >10%. Nevertheless, when applying Equation (14) assuming 176 participants, the MMOE is 1.12, and thus the error is expected to be 12%, only just over the 10% recommendation. Hence, this existing dataset appears to have a reasonable sample size for external validation, which would have been useful for Hudda et al to know at the time.

6 | DISCUSSION

We have proposed closed-form sample size calculations for studies externally validating a prediction model for a continuous outcome. These aim to ensure the sample size is large enough to precisely estimate key measures of predictive performance (R^2, CITL, and calibration slope) and the residual variances in calibration models. This led to four criteria, and the largest sample size required satisfying all four criteria is the recommended minimum sample size needed in the external validation dataset. Our work builds on minimum sample size calculations for model development.[4,38]

As with any sample size calculation, assumptions are required to implement our proposed approach. In particular, researchers must specify the model's anticipated R^2_{val}, $\widehat{\text{var}}(Y_i)$, and $\widehat{\lambda}_{\text{cal}}$ in the validation dataset. As discussed, a simple starting point is to assume these will be the same as those reported for the original model development study, especially if the target population (for validation) is similar to that in the model development study. Then the researcher might consider sample sizes based on slight adjustments; in particular, assuming the model may perform slightly worse than in the development dataset. Our example illustrated this for a prediction model of fat-free mass in children, where we assumed an R^2_{val} of 0.8 rather than the 0.90 or 0.95 values reported in the original model development study. Lower values may be even more important to consider in situations where the development dataset was small (such that reported performance statistics were estimated with large uncertainty); the developed prediction model did not adjust for overfitting using, for example, penalization and shrinkage techniques (such that reported performance statistics are likely to be optimistic); and in situations where the intention is to validate the model in a different population or setting from that used in the development study. Larger sample sizes may be needed if missing data are expected, and if a model's predictive performance in key subgroups (eg, males, females) is of interest.

Section 5 discussed how to use our calculations when an existing dataset (of a fixed sample size) is already available, in order to gauge the expected precision of estimates conditional on the sample size available. Ideally the dataset will be large enough to ensure precise estimates, as then more robust conclusions about predictive performance will be possible. However, we recognize that even when datasets are small, obtaining estimates of predictive performance is still useful; in particular, these could ultimately be combined in a meta-analysis.[39] It is important that datasets for external validation are high quality and applicable to the target population, setting, and timing of implementing the prediction model in practice. Adequate sample size does not overcome issues in quality and applicability.[39-41]

We chose to focus on R^2, CITL, and calibration slope as these are key performance measures; ensuring precise estimation of residual variances is also important, as they are used to calculate the aforementioned predictive performance measures and also mean-squared error. We anticipate that the largest sample size will usually be driven by criterion (i), (iii), or (iv). Further work might consider precise estimation of calibration curves,[11,22,42] and extension to non-continuous outcomes is needed, building on work of others.[11,17,43] Closed-form sample size solutions are transparent and quick to implement, but more difficult to derive for binary and time-to-event outcomes. Jinks et al do suggest closed-form sample size calculations for precisely estimating the D statistic for time-to-event prediction models.[44] Also, we only focused on statistical measures of predictive performance, and not on clinical utility or impact of using the model to inform healthcare decisions (eg, initiation of treatment).

Finally, sometimes the sample size for an external validation dataset must also be large enough for model updating, for example, when the researcher aims to recalibrate one or a few of the model parameters to the target population of interest. Then, the required sample size needs to meet the criteria described in this article (for external validation), and also those criteria proposed for model development (as model updating is akin to model development[5]). The exact sample size needed for model updating depends on how the model is to be updated (eg, which parameters, and indeed how many parameters, are to be revised) and whether additional predictors are to be included. Riley et al provide advice for this and other model development situations.[5]

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their constructive feedback that helped improve our article upon revision. We are grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. Lucinda Archer is supported by funding from the European Horizon 2020 Research and Innovation Programme under grant agreement No 777090. Kym Snell is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR Launching Fellowship). Joie Ensor is funded by NIHR Clinical Trials Unit Support Funding, Supporting Efficient/Innovative Delivery of NIHR Research. Mohammed Hudda is supported by a British Heart Foundation PhD Studentship (FS/17/76/33286). Gary Collins is

supported by the NIHR Biomedical Research Centre, Oxford and Cancer Research UK programme grant (C49297/A27294). Lucinda Archer, Richard Riley and Kym Snell are supported by funding from the Evidence Synthesis Working Group, which is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR) [Project Number 390]. The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for the ALSPAC study.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ETHICS STATEMENTS

Ethical approval for the ALSPAC study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. The views expressed are those of the authors and not necessarily those of the BHF, Cancer Research UK, the NHS, the NIHR, the Department of Health or the EU.

ORCID

Lucinda Archer https://orcid.org/0000-0003-2504-2613
Kym I. E. Snell https://orcid.org/0000-0001-9373-6591
Joie Ensor https://orcid.org/0000-0001-7481-0282
Mohammed T. Hudda https://orcid.org/0000-0001-7894-1159
Gary S. Collins https://orcid.org/0000-0002-2772-2316
Richard D. Riley https://orcid.org/0000-0001-8699-0735

REFERENCES

1. Riley RD, van der Windt D, Croft P, Moons KG, eds. *Prognosis Research in Healthcare: Concepts, Methods and Impact*. Oxford, UK: Oxford University Press; 2019.
2. Harrell FE Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Second ed. New York: Springer; 2015.
3. Steyerberg EW. *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating*. New York: Springer; 2009.
4. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I - continuous outcomes. *Stat Med*. 2019;38(7):1262-1275. https://doi.org/10.1002/sim.7993.
5. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
6. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515-524. https://doi.org/10.7326/0003-4819-130-6-199903160-00016.
8. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med*. 2006;144(3):201-209.
9. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61(11):1085-1094. https://doi.org/10.1016/j.jclinepi.2008.04.008.
10. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68(3):279-289. https://doi.org/10.1016/j.jclinepi.2014.06.018.
11. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176. https://doi.org/10.1016/j.jclinepi.2015.12.005.
12. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13:33. https://doi.org/10.1186/1471-2288-13-33.
13. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in, prediction research: a clinical example. *J Clin Epidemiol*. 2003;56(9):826-832. https://doi.org/10.1016/S0895-4356(03)00207-5.
14. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245-247. https://doi.org/10.1016/j.jclinepi.2015.04.005.
15. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ*. 2010;340:c2442. https://doi.org/10.1136/bmj.c2442.
16. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40. https://doi.org/10.1186/1471-2288-14-40.
17. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016;35(2):214-226. https://doi.org/10.1002/sim.6787.
18. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140.
19. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381. https://doi.org/10.1371/journal.pmed.1001381.

20. Hudda MT, Fewtrell MS, Haroun D, et al. Development and validation of a prediction model for fat mass in children and adolescents: meta-analysis using individual participant data. *BMJ*. 2019;366:l4293. https://doi.org/10.1136/bmj.l4293.
21. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. https://doi.org/10.1186/s12916-019-1466-7.
22. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517-535. https://doi.org/10.1002/sim.5941.
23. Copas JB. Regression, prediction and shrinkage. *J R Stat Soc B Methodol*. 1983;45(3):311-354.
24. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res*. 1997;6(2):167-183. https://doi.org/10.1177/096228029700600206.
25. Stein C. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. *Proc Third Berkeley Symp Math Stat Prob*. 1956;1:197-206.
26. Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica*. 2001;55:17-34.
27. Kirchner J. Data Analysis Toolkit #10: Simple linear regression. http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf. 1996
28. Tan L. Confidence Intervals for Comparison of the Squared Multiple Correlation Coefficients of Non-nested Models. Electronic Thesis and Dissertation Repository (Paper 384). 2012
29. Wishart J. The mean and second moment coefficient of the multiple correlation coefficient in samples from a normal population. *Biometrika*. 1931;22:353-361.
30. Lee YS. Tables of the upper percentage points of the multiple correlation. *Biometrika*. 1971;59:175-189.
31. Kelley K. Confidence intervals for standardized effect sizes: theory, application, and implementation. *J Stat Softw*. 2007;20(8):24. https://doi.org/10.18637/jss.v020.i08.
32. Kelley K. Methods for the behavioral, educational, and social sciences: an R package. *Behav Res Methods*. 2007;39(4):979-984. https://doi.org/10.3758/bf03192993.
33. Kelley K. MBESS (Version 4.0.0 and higher) [computer software and manual]. https://CRAN.R-project.org/package=MBESS. 2017.
34. Montgomery DC, Peck EA, Vining GG. *Introduction to Linear Regression Analysis*. Third ed. New York: Wiley; 2001.
35. Boyd A, Golding J, Macleod J, et al. Cohort profile: the 'children of the 90s'—the index offspring of the Avon longitudinal study of parents and children. *Int J Epidemiol*. 2013;42(1):111-127. https://doi.org/10.1093/ije/dys064.
36. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort profile: the Avon longitudinal study of parents and children: ALSPAC mothers cohort. *Int J Epidemiol*. 2013;42(1):97-110. https://doi.org/10.1093/ije/dys066.
37. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol*. 2014;14(1):135. https://doi.org/10.1186/1471-2288-14-135.
38. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296. https://doi.org/10.1002/sim.7992.
39. Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460. https://doi.org/10.1136/bmj.i6460.
40. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. https://doi.org/10.7326/M18-1377.
41. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. https://doi.org/10.7326/M18-1376.
42. Austin PC, Steyerberg EW. Bootstrap confidence intervals for loess-based calibration curves. *Stat Med*. 2014;33(15):2699-2700. https://doi.org/10.1002/sim.6167.
43. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol*. 2003;56(5):441-447.
44. Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Med Res Methodol*. 2015;15:82. https://doi.org/10.1186/s12874-015-0078-y.

How to cite this article: Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine*. 2020;1–14. https://doi.org/10.1002/sim.8766